Analysis plan for “A multi-sensory tutoring program for students at-risk of reading difficulties in kindergarten and first grade: Evidence from a randomized field experiment”

Our primary objective is to evaluate the short-run effects of the program. We present this analysis first below together with planned sensitivity analyses. Then we describe how we test hypotheses related the timing of getting the program, i.e., if there is a difference between the treatment and the waitlist control after the latter have also received the program. Lastly, we describe exploratory analyses. All statistical analyses will be conducted by the first two authors (Bøg and Dietrichson).

Short-run effects of the program

We will test hypotheses pertaining to the short-run effects of the program in a linear regression framework where we include a treatment indicator as an independent variable in the regression together with baseline test scores, and strata fixed effects (indicators for matched pairs/triples) as covariates.

Most students are given the program one-to-one, but some students receive the program in small groups. Weiss et al. (2016) show that models with a random group effect or a using cluster-robust variance estimators may overestimate variance in such cases. We will use standard errors that are robust to heterogeneity of unknown form in our primary analysis and test whether the results are sensitive to this choice by cluster the standard errors on the groups (using the cluster-option in Stata).

Our primary analysis will use the sample with pre-treatment test scores (every randomized student has taken each pre-test) and at least one post-treatment score. That is, treatment in this case is not defined by having received training, only on being randomized to treatment. The estimate is therefore of an intention-to-treat (ITT) type. We describe how we will examine the effect of being more or less exposed to training in the section about exploratory analyses below.

Our hypotheses (H) about the short-run effects of the program will be tested using the contrast between the treatment group and the waitlist control group on the six outcome measures after the treatment group but before the waitlist control group has received the program. That is, H1-H6 corresponds to testing whether the coefficient on the treatment indicator is different from zero using the following tests as outcome variables:

1. The decoding test
2. The letter knowledge test
3. The phonological awareness test
4. How easy or hard do you think it will be to learn how to read?
5. How much fun do you think it will be to learn how to read?
6. How much would you like to learn how to read?

Testing multiple hypotheses risks increasing the familywise error rate (FWER), i.e., the risk of rejecting at least one true null hypothesis, above the desired level of significance (Romano & Wolf, 2005). We will use a type of sequentially rejective multiple test procedure (see e.g., Bretz et al., 2009), designed to achieve strong control of the FWER, set to α = 0.05. Strong control implies that the FWER is not bigger than α for any configuration of true and null hypotheses (Romano & Wolf, 2005).

We have two primary outcomes, corresponding to H1 and H2. H1 is our most important hypothesis, as the main objective of the program is to teach students to decode simple words. We will test H1 first, using a significance level of α = 0.05. This is our first testing level (or “family” in the terminology of Bretz et al., 2009). We let H2 be the second testing level. That is, if H1 is rejected with $p < 0.05$, then we will test H2 using α = 0.05. The fixed testing sequence implies that if H1 cannot be rejected (i.e., $p > 0.05$), then we cannot formally reject H2, regardless of its *p*-value.

Our third testing level is H3-H6, corresponding to our secondary outcomes. H3-H6 are more difficult to order in terms of importance. The phonological awareness test is closer to the content of the program, but has drawbacks in terms of e.g., being researcher-developed (see e.g., Cheung & Slavin, 2016, for problems with researcher-developed having larger effect sizes) and potential ceiling effects, as discussed above. H4-H6 are furthermore difficult to order, as they are testing different aspects of motivation, enjoyment, and self-efficacy. For these reasons, we have no pre-defined testing hierarchy regarding H3-H6. Instead, we use a stepdown procedure (e.g., Heckman et al., 2010; Romano & Wolf, 2005), the first step of which is to jointly test if H3-H6 can be rejected. (Again, the fixed testing sequence between levels implies that we can only formally reject any of H3-H6 if we first reject H2 and therefore, if H1 was also rejected).

Stepdown methods use test statistics for individual tests to test joint hypotheses. To control the FWER, the individual tests can only use a fraction of α, denoted α_i, $i \in \{3,4,5,6\}$. Let the initial allocation be equal among the hypotheses, i.e., $\alpha_3 = \ldots = \alpha_6 = \alpha/4 = 0.0125$. If $p > 0.0125$ for all hypotheses, the null is not rejected for any of them. If any hypothesis is rejected (i.e., $p < 0.0125$), consider the most significant hypothesis (lowest p-value) rejected, remove it from the set, and distribute allocation for this hypothesis fully and equally among the remaining hypotheses (Bretz et al., 2009). The remaining three hypotheses are then tested using $\alpha_i = \alpha/3 = 0.0167$ as the significance level. Again, we stop if no hypothesis can be rejected, and otherwise continue distributing the allocation and testing until no hypothesis can be rejected, or there are no hypotheses left to test.

Sensitivity analyses. We will report descriptive statistics over pre-treatment test scores and student characteristics to examine whether the randomization produced balanced treatment and control groups. Beside standard balancing tests, comparisons between specifications with and without covariates may indicate that the randomization did not create balanced treatment and control groups, and we will report specifications without covariates.

To examine sensitivity to missing observations, we will drop all pairs/triples with missing outcomes, thus balancing treatment and control groups regarding attrition. If this procedure yields significantly different results compared to the baseline (point estimates outside the 95% confidence interval), we will examine whether attrition and missing observations on dependent variables vary over treatment and control groups using logistic regressions. If we do not find significant differences, we will rely on the primary specifications for our conclusions. In case of significant differences, we will base our conclusions on the sample with dropped pairs/triples and consider various methods of correction, e.g., by inverse probability weighting (see e.g., Conti et al. 2016).

To assess whether our results are sensitive to the assumptions made on the distribution of the standard errors when using regular inference methods, we will use randomization (or permutation based) inference methods, as described by e.g., Young (2016) and Athey & Imbens (2017). If the outcomes of the randomization tests differ from those obtained by regular inference methods, we will base our conclusions on the randomization tests if Stata's test of normality rejects the null of normally distributed errors on the 5 percent level. We will use the same testing procedure to control for multiple hypothesis testing, which also achieve strong control of the FWER conditional on i) that the same draw of permutation is used to compute the all test statistics at each stage, and ii) that the permutation set from which permutations are drawn is chosen such that, under the null hypotheses, the distribution of the data is invariant for each permutation (Heckman et al., 2010; the latter condition is needed more in in general for randomization tests to be valid).

Lastly, depending on how many words students can decode, there may be a small number of words in the decoding test that overlaps with the material used in the second step of the program. If students in the treatment group learn these by heart, the test may overestimate their decoding skills, and we will test whether our results are sensitive to removing points from the overlapping words.

Effects of program timing

We will test hypotheses about the effects of program timing – whether it matters to have been waitlisted – in the same regression framework as described above. Outcome variables will be the measures used to evaluate the short-run effects, but measured at follow-up.

Sensitivity analysis. We will perform the same sensitivity analyses as mentioned for the short-run effects.

Exploratory analyses

Exposure and imperfect compliance. Some students may get fewer sessions than intended due to for example implementation problems at schools. We will measure the number of sessions given to each student. If there are differences in exposure (or imperfect compliance), we will examine the consequences for the short-run effects. As exposure and compliance may be endogenous to treatment, we will estimate the effects using instrumental variables (IV), as suggested by e.g., Angrist et al. (1996). The variable indicating the assignment to treatment and waitlist control groups will be the instrument for the number of sessions (possibly defined as an indicator, depending on the distribution of sessions) in a two-stage least squares (2SLS) estimation. We will also try to increase the instrument strength by creating new instruments from interactions between the treatment indicator and pre-determined variables (primarily baseline test scores). We will use the combination of instruments with the highest first stage F-value, and only try the 2SLS if we find a first stage F-value above 10.

Moderator and heterogeneity analyses. If there are short-run effects, we will explore heterogeneity of these effects over baseline measures. As we expect most students' baseline decoding scores to be very close to zero, and the letter knowledge and phonological awareness to be highly correlated, we limit the use of test score moderators to the letter knowledge test. We similarly expect the three questions pertaining to motivation, enjoyment, and self-efficacy to be highly correlated, and will use the unweighted average score. For both moderators, we create a value indicating whether a student is above the median and interact this indicator with the treatment indicator. Both moderators will be included in the same regression. Two schools have prior experience with implementing the program (for earlier cohorts), and, as mentioned, some schools may choose to participate with a second cohort. Implementing teachers in these schools thus have more experience with the program. We will test if the results are heterogeneous over prior experience by interacting the treatment indicator with an indicator for students in these schools.

Cost-effectiveness

To evaluate costs we collect the following information: costs of developing the intervention, opportunity costs of training tutors, and the costs of delivering the intervention. An important short term benefit lies in the avoidance of more costly services that this group of students might otherwise require. We will interview school personnel and search for historical school records to establish what, if any, interventions students in our target group would have received in the first years of school, if they had not been given Läsklar.

References mentioned in registration or analysis plan

Angrist, J. D., Imbens, G. W., & Rubin, D. B. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434), 444-455.

Athey, S. & Imbens, G. (2017). The econometrics of randomized experiments. In: Banerjee, A. & Duflo, E. (Eds.), *Handbook of Field Experiments, Volume 1*. Amsterdam: Elsevier.

Bretz, F., Maurer, W., Brannath, W., & Posch, M. (2009). A graphical approach to sequentially rejective multiple test procedures. *Statistics in Medicine*, 28, p. 586-604.

Bøg, M., Dietrichson, J., & Isaksson, A. A. (2017). *A multi-sensory literacy program for at-risk students in kindergarten – Promising results from a small-scale Swedish intervention*. Unpublished manuscript.

Cheung, A., & Slavin, R. (2016). How methodological features affect effect sizes in education. *Educational Researcher*, 45(5), 283-292.

Conti, G., Heckman, J. J., & Pinto, R. (2016). The effect of two influential early childhood interventions on health and healthy behavior. *Economic Journal*, 126, F28-F65.

Elwér, Å., Fridolfsson, I., Samuelsson, S., & Wiklund, C. (2016). *LäSt – Test i läsförståelse, läsning och stavning för åk 1-6*. Hogrefe Psykologiförlaget: Stockholm.

Heckman, J. J., Moon, S. H., Pinto, R., Savelyev, P., & Yavitz, A. (2010). Analyzing social experiments as implemented: A reexamination of the evidence from the HighScope Perry Preschool Program. *Quantitative Economics*, 1, 1-46.

Johansson, M-G. (2009). *LäsEttan*. Stockholm: Natur & Kultur.

Romano, J. P., & Wolf, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*, 100(469), 94-108.

Weiss, M. J., Lockwood, J. R., & McCaffrey, D. F. (2016). Estimating the standard error of the impact estimator in individually randomized trials with clustering. *Journal of Research on Educational Effectiveness*, 9(3), 421-444.

Young, A. (2016). *Channelling Fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results*. Unpublished manuscript.